DOCUMENT RESUME

ED 230 566 TM 830 234

AUTHOR	Dorans, Neil J.; Kulick, Edward
TITLE	Assessing Unexpected Differential Item Performance of Female Candidates on SAT and TSWE Forms Administered in December 1977: An Application of the Standardization Approach.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-83-9
PUB DATE	Feb 83
NOTE	54p.; Some tables may be marginally legible due to small print.
AVAILABLE FROM	Educational Testing Service, Research Publications, R116, Princeton, New Jersey, 08541.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Ability; College Entrance Examinations; *Females; High Schools; *Item Analysis; Language Tests; Latent Trait Theory; Models; Standardized Tests; *Statistical Analysis; *Test Bias; Test Items
IDENTIFIERS	*Scholastic Aptitude Test; Standardization; *Test of Standard Written English

ABSTRACT

A new approach to assessing unexpected differential item performance (item bias or item fairness) was developed and applied to the item responses of males and females to Scholastic Aptitude Test and Test of Standard Written English items administered operationally in December 1977. While the main body of the report describes the particulars of the present application and delineates the essential features of the approach, a technical appendix describes the standardization approach in detail. The primary goal of the standardization approach is to control for differences in subpopulation ability before making comparisons between subpopulation performance on test items. By so doing, it removes the contaminating effects of ability differences from the assessment of item fairness. Of the total of 195 items studied, the standardization approach identified only a handful as meriting careful review for possible content bias. Of these few, only one item exhibited a clearly unacceptable degree of unexpected differential item performance between males and females that could be attributed to content bias. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-83-9

ED230566

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ASSESSING UNEXPECTED DIFFERENTIAL ITEM PERFORMANCE OF FEMALE CANDIDATES ON SAT AND TSWE FORMS ADMINISTERED IN DECEMBER 1977: AN APPLICATION OF THE STANDARDIZATION APPROACH

Neil J. Dorans
Edward Kulick

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. C. Weidenmiller
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

February 1983

TM 830 234

Educational Testing Service
Princeton, New Jersey

ASSESSING UNEXPECTED DIFFERENTIAL ITEM PERFORMANCE OF FEMALE CANDIDATES ON SAT AND TSWE FORMS ADMINISTERED IN DECEMBER 1977: AN APPLICATION OF THE STANDARDIZATION APPROACH[1]

Neil J. Dorans
Edward Kulick

Educational Testing Service

February, 1983

[1]This approach developed as an outcome of several conversations with Paul H. Holland about the various problems associated with assessing unexpected differential item performance. The authors are grateful for his helpful comments. This approach, which is still evolving, has also been shaped by the comments and suggestions of Thomas F. Donlon, Gary L. Marco and Nancy S. Petersen. We also thank Lawrence J. Stricker for his careful review of an earlier draft. Finally, without the programming and systems development work of Edwin O. Blew and Karen Carroll, this approach would have remained a system of equations.

Copyright © 1983. Educational Testing Service. All rights reserved.

Abstract

A new approach to assessing unexpected differential item performance (item bias or item fairness) is developed and applied to the item responses of males and females to SAT/TSWE items administered operationally in December 1977. While the main body of the report describes the particulars of the present application and delineates the essential features of the approach, a technical appendix describes the standardization approach in detail. The primary goal of the standardization approach is to control for differences in subpopulation ability before making comparisons between subpopulation performance on test items. By so doing, it removes the contaminating effects of ability differences from the assessment of item fairness. Of the total of 195 items studied, the standardization approach identified only a handful as meriting careful review for possible content bias. Of these few, only one item exhibited a clearly unacceptable degree of unexpected differential item performance between males and females that could be attributed to content bias.

ASSESSING UNEXPECTED DIFFERENTIAL ITEM PERFORMANCE OF FEMALE CANDIDATES ON SAT AND TSWE FORMS ADMINISTERED IN DECEMBER 1977: AN APPLICATION OF THE STANDARDIZATION APPROACH

Those who develop and review the Scholastic Aptitude Test (SAT) are aware of the diversity of the test-taking population and attempt to construct tests based on a broad sampling of tasks and topics that tend not to favor any subgroup of the population. Donlon (1981) discussed the checks that are performed on the SAT to guard against favoritism towards any subgroup. In that article, Donlon summarized procedures used in the test development process to ensure that items or test questions are appropriate for various subgroups as well as the types of statistical checks performed to evaluate item appropriateness.

Carlton and Marco (1982), in a review of methods used at Educational Testing Service to detect and eliminate possible favoritism in items, discussed several studies that have examined performance on SAT items across different subpopulations. Included in their review were six studies that were conducted to monitor differential item performance of various groups on several forms of the SAT and its companion test, the Test of Standard Written English (TSWE). The purposes of this monitoring are:

(1) to ensure that the SAT and TSWE remain appropriate over time for major subgroups of the SAT candidate population, and

(2) to identify possible content factors related to differential item performance that would help test developers construct fair tests.

Dorans (1982) reviewed the five of those six studies that examined Black/White candidate performance on SAT/TSWE items from forms of the SAT/TSWE that have the current content and format specifications. In the present report, the statistical method of standardization is used to examine whether there are

unexpected differences in item performance across different subpopulations of the Scholastic Aptitude Test test-taking population.

Unexpected Differential Item Performance

Unexpected differential item performance exists when there are differences in item performance that cannot be accounted for by differences in subgroup ability. An item is exhibiting unexpected differential item performance when the expected performance on the item is lower for examinees from one group than for examinees of equal ability from another group or other groups. If we let S represent ability as measured by total score[1] on the standard College Board 200-to-800 SAT scale (or on the 20-to-60 TSWE scale), and X represent an item score (1 if the answer to the question is correct and 0 if the answer is incorrect), then an item is free of unexpected differential item performance when it satisfies the following equality

$$(1) \qquad P_g(X=1|S) = P_{g'}(X=1|S) \qquad \text{for all subpopulations } g \text{ and } g',$$

where $P_g(X=1|S)$ is defined as the probability that candidates from subpopulation g who have total test scores equal to S will answer the item correctly. For example, if male and female candidates with the same total test scores do not

---

[1] It is recognized that use of reported scaled score as the control variable can be criticized because it is not a perfect measure of ability and because it is an internal criterion, i.e., performance on an item is related to total score performance in part because that item went into the determination of total score. Nonetheless, reported scaled score is probably the best control variable available for studies of unexpected differential item performance.

have equal probabilities of successful performance on an item, this difference is taken as evidence of unexpected differential item performance for male and female candidates at that particular score level. Note that lack of unexpected differential item performance does not imply that there are no differences in item performance across subgroups of the Scholastic Aptitude Test candidate population. Unexpected differential item performance does not refer to differences in overall subgroup performance on an item but rather to differences in conditional item performance where the requisite condition before comparison is identical total test score.

Several methods have been suggested for identifying unexpected differential item performance, or item bias as it is frequently referred to in the literature. The handbook by Berk (1982) attests to this fact. For a single comprehensive review of the more popular methods, including the transformed item difficulty or delta-plot method, item response theory methods and chi-square approaches see Shepard, Camilli and Averill (1981). Most of these methods, however, have exhibited undesirable sensitivities to differences in overall subpopulation ability or differences in item quality (discrimination). Two of these methods (transformed item difficulty and a chi-square approach) were employed in earlier studies of the Scholastic Aptitude Test that were reviewed by Dorans (1982). Both methods are subject to misclassifying items as unfair towards a particular subgroup because of methodological sensitivities to differences in subpopulation ability. The methodology employed in the current study controls for differences in subpopulation ability through the statistical method of standardization.

Standardization is a technical term that, unfortunately, has more than one meaning. In one usage, standardization typically refers to a numerical operation which transforms a set of numbers with a particular mean (average score)

and standard deviation (spread of scores about the average score) to a set of numbers that has a certain "standard" mean and standard deviation. This is not the meaning of standardization as used in this report.

Rather, we shall use standardization to mean that one variable is standardized with respect to some other variable before making comparisons between groups. This type of standardization enables one to control for differences in subpopulation ability while making comparisons of the performance of these subpopulations on items. The procedures used in this study require a very large data base in order to ensure the stability of the conditional probabilities obtained at each score level in each subpopulation under investigation. Fortunately, there are large data bases available for the Scholastic Aptitude Test. Other methods of standardization may be used with smaller sample sizes, e.g. Alderman and Holland (1981). A general approach to assessing unexpected differences in item performance via standardization is described in detail in the appendix, where a mathematical formulation is presented and the method's similarities to and differences from the item response theory approach is discussed.

Standardization

In this section, the essential features of standardization are described. The conditional probability of successful performance on an item, $P_g(X=1|S)$, is the raw datum for the standardization method. For each score level S, there is a conditional probability of successful performance. Studies of unexpected differential item performance focus on differences in conditional probability of successful item performance between a study group and a base group. In this

first study, female SAT candidates are the study group, while male candidates are the base group.

Figure 1 contains plots of the conditional probability of successful performance for both males and females on an analogy item appearing on Form ZSA5. Male conditional percent corrects are denoted by squares (□) at each score level, while female conditional percent corrects are denoted by asterisks (*). (Note that there are no asterisks at scaled scores of 770 and 800, which indicates there were no females at those two scaled score levels.) In this particular figure, the asterisks and squares tend to lie on top of one another. This consistent and high degree of overlap is evident in Figure 2, which is a plot of differences in conditional probabilities for this item. Note that almost all the asterisks in Figure 2 lie very close to the line of zero difference. This particular analogy item exhibits very little unexpected differential item performance.

The analogy item portrayed in Figures 3 and 4 serves as a striking contrast to that depicted in Figures 1 and 2. Here, the squares (males) are higher than the asterisks (females) at almost every scaled score level. In fact, between scaled scores of 250 and 500, the difference between the female conditional probabilities and the male conditional probabilities tends to be .2, i.e., the probability that a male with a given scaled score in that range will answer that analogy item correctly exceeds the probability that a female with the same exact scaled score will answer the item correctly by the substantial amount of .2. Clearly, this particular item exhibits a substantial amount of unexpected differential item performance.

Examination of conditional probability plots such as those depicted in Figures 1 and 3 and difference plots like those in Figures 2 and 4 enables

Conditional Probability of Successful Item Performance
for Both Males and Females on Two
Verbal Items from SAT Form ZSA5

Item 39

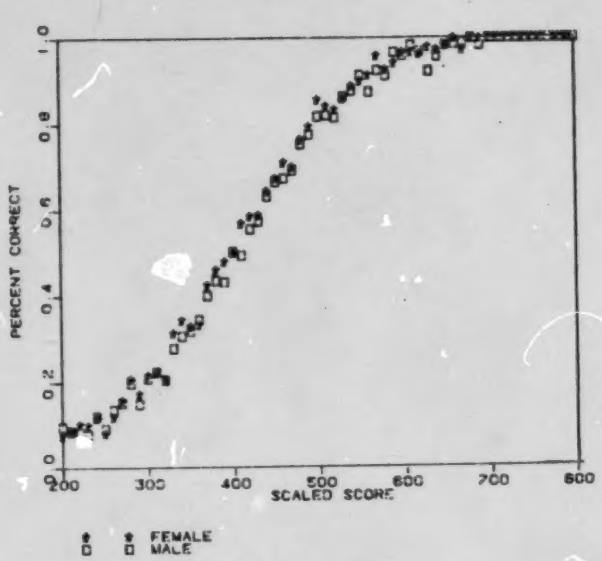


Figure 1

Item 63

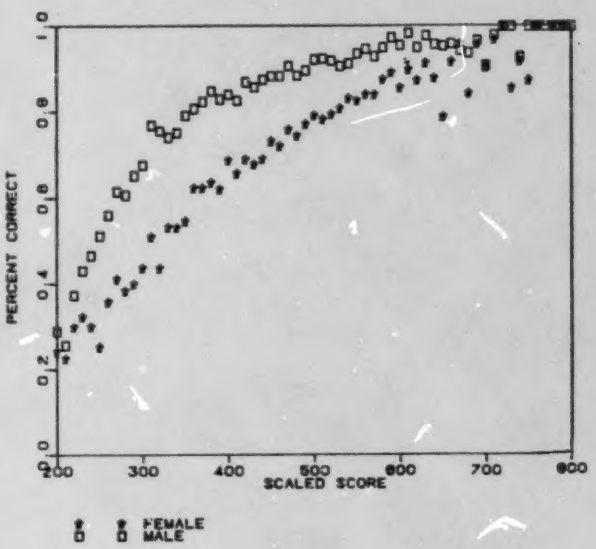


Figure 3

Difference Plots of Two Verbal Items from SAT Form ZSA5

Item 39

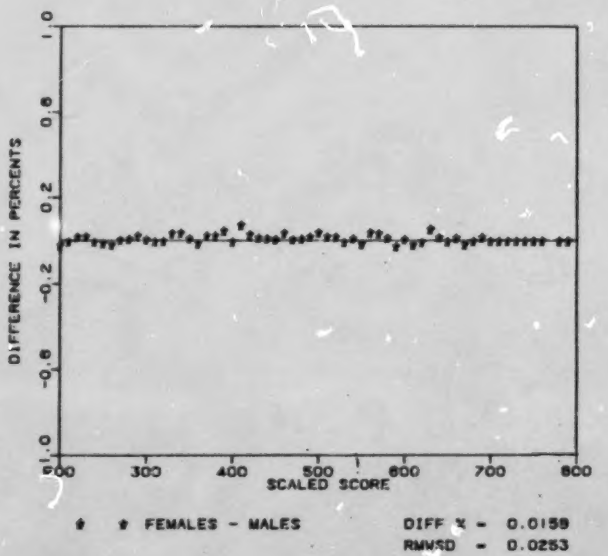


Figure 2

Item 63

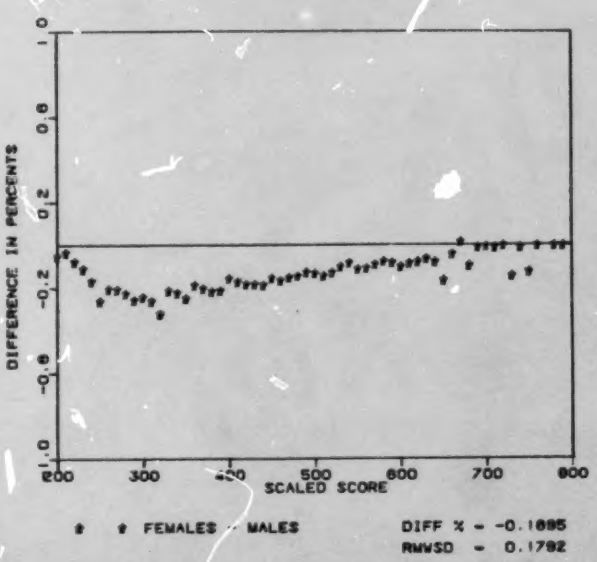


Figure 4

one to look for evidence of unexpected differential item performance at fixed score levels. In effect, the plots allow one to control for ability before comparing item performance across subpopulations. Consequently, for each item there is potential for unexpected differential item performance that can be summarized via some numerical index. One such index is the difference in conditional probabilities of successful performance at that score level. If there are 61 observed score levels, such as there are on the College Board SAT scale that ranges from 200-to-800 in steps of 10, then there are 61 such differences for each item. Clearly there exists a need for an economical summary of these differences. Standardization provides that summary.

The application of the standardization procedure, in which the marginal ability distribution of the female standardization group serves as a weighting function, yields several summary indices of item performance. First, there is the observed percent correct P_f for the female study group obtained by taking a weighted sum of the 61 conditional probabilities of successful performance observed in the female study group, where the relative frequencies at each of the 61 scaled score levels in the female study group serve as the weights. These same weights are applied to the 61 conditional probabilities observed in the male base group to produce an index of expected item performance $\hat{P}_f$ for the female study group. The difference between P_f and $\hat{P}_f$, $D_f = P_f - \hat{P}_f$, is one index of unexpected differential item performance. If there is no unexpected differential item performance, D_f should equal zero. A positive D_f indicates that the study group exceeds its expected performance, while a negative D_f indicates that the item is harder than expected for the study group. Since D_f is a signed index, it is insensitive to crossovers in the conditional success distributions of the base and study groups. An unsigned discrepancy

index that can be used with D_f is the root mean weighted squared difference ($RMWSD_f$). The $RMWSD_f$ for an item is obtained by weighting each difference in conditional probabilities of successful item performance between the study and base groups by that difference (which is equivalent to squaring the difference) and by the frequency of scores in the female standardization group at each scale score level, summing this weighted difference across the 61 scaled score levels, dividing this sum by the number of candidates in the standardization group, and taking the square root of the result. The mathematical formula for the $RMWSD_f$ is

$$(2) \qquad RMWSD_f = \left(\sum_{s=1}^{S} N_{fs+}(P_{fs} - \hat{P}_{fs})^2 \Big/ \sum_{s=1}^{S} N_{fs+} \right)^{1/2}$$

where S is the number of score levels, N_{fs+} is the number of individuals at score level s in subpopulation f, P_{fs} is the conditional probability of successful performance in subpopulation f at score level s, and $\hat{P}_{fs}$ is the predicted value of P_{fs}. Note that typically $\hat{P}_{fs} = P_{bs}$, where P_{bs} is the conditional probability of successful performance observed at score level s in the male base group. Given the definition of D_f as

$$(3) \qquad D_f = \sum_{s=1}^{S} N_{fs+}(P_{fs} - \hat{P}_{fs}) \Big/ \sum_{s=1}^{S} N_{fs+}$$

it can be shown that

$$(4) \qquad RMWSD_f = \left(D_f^2 + \sum_{s=1}^{S} N_{fs+}(D_{fs} - D_f)^2 \Big/ \sum_{s=1}^{S} N_{fs+}\right)^{1/2}$$

where $D_{fs} = P_{fs} - \hat{P}_{fs}$. Since, this index is unsigned, any difference produces a positive discrepancy. Consequently, every item will have a positive $RMWSD_f$. An item exhibiting substantial unexpected differential item performance will have a large $RMWSD_f$.

Equation (4) expresses $RMWSD_f$ as the square root of two additive components, the square of a constant directional discrepancy, which is D_f^2, and an index of residual crossover, i.e., a sum of weighted squared differences in conditional probabilities after adjusting for the constant difference, which is the second component in (4). While the D_f^2 portion is probably systematic and indicative of unexpected differential item performance, the residual crossover component may or may not be indicative of systematic unexpected differential item performance because it does not allow random differences to cancel out. As such, the primary purpose of the residual crossover component is to flag an item for closer examination.

A problem faced by any investigation which seeks to detect and quantify unexpected differential item performance, regardless of methodology, is the determination of what level of unexpected differential item performance should evoke concern. One could argue that any difference should evoke concern. This, however, would be an extreme position that ignores the fact that measurement systems are always contaminted by noise. In the present study, we examined distributions of root mean weighted squared differences ($RMWSD_i$) to empirically determine a cutoff point which defines a substantial amount of unexpected differential item performance. Examination of these frequency distributions led us to conclude that an item with a $RMWSD_f$ greater than or equal to .08 merits careful investigation, while an item with a $RMWSD_f$ less than .08 does not

require additional study. As we acquire more experience with applying the standardization approach to other data bases, a better cutoff may evolve.

In combination, D_f and $RMWSD_f$ provide a statistical description of an item that will enable us to ascertain the degree of unexpected differential item performance obtained in the female study group.

Test Form and Sample Used in This Study

SAT Form ZSA5 and TSWE Form E11, administered in December 1977, were used in this study. Stern (1977) previously described the psychometric properties of TSWE Form E11; and Cook and Nutkowitz (1979), the psychometric properties of SAT Form ZSA5. Since the psychometric properties of ZSA5/E11 are described in detail in the test analysis reports just cited, only the most salient characteristics are summarized here. Both the verbal and mathematical sections of Form ZSA5 had fairly typical reliabilities (and scaled score standard errors of measurement) of .914 (32) and .916 (33), respectively, in a spaced sample of 1,895 candidates from the total group of 166,311 candidates who took Form ZSA5 in December, 1977. The mean equated delta, an index of test difficulty described by Hecht and Swineford (1981) and Walker (1981), for the verbal section was 11.3, which indicated the test was slightly easier than intended. For the mathematical section, the mean equated delta was 12.4, slightly more difficult than intended. TSWE Form E11 had a fairly typical reliability of .887 in a spaced sample of 1,615 candidates from the total group of 84,144 who took TSWE Form E11 in June 1976. The mean equated delta was 9.3, slightly easier than intended.

The basic data for this study were the item responses of 21,835 male candidates and 21,209 female candidates who took the 85 verbal, 60 mathematical and 50 TSWE items that appeared in the operational sections of the Forms ZSA5 and E11 that were administered in December, 1977. The combined sample of 43,044 was representative of the total group that took ZSA5/E11 at that administration.

Procedure

The focus of the present study is on the assessment of unexpected differential item performance for female candidates on Forms ZSA5 and E11 items. In this particular application of the general standardization technique, the study group is the female candidate subpopulation. The standardization group supplies the standard ability distribution used by the standardization approach. Any subgroup including a composite group or a hypothetical group can be used as the standardization group. Since the standard ability distribution serves as a weighting function, it is advisable to use each study group as its own standardization group thereby enabling use of a weighting function that mirrors the relative frequency at each score level in the study group. The male candidate subpopulation, as the majority group, was chosen as the base group, i.e., the subpopulation that supplies the model for item performance as a function of ability. The model is the conditional probability of successful performance on the item given ability. The largest subpopulation was used as the base group in order to produce the most statistically stable model of item performance given test score that can be attained. Table 1 contains the marginal score distribution for the female study group and male base group for SAT-Verbal, SAT-Mathematical, and TSWE. Note that the largest weights (relative frequency in

Table 1

Frequency Distributions and Summary Statistics of Males' and Females' Verbal, Mathematical, and TSWE Scaled Scores

Scaled Score	VERBAL Male f	VERBAL Male % below	VERBAL Female f	VERBAL Female % below	MATHEMATICAL Male f	MATHEMATICAL Male % below	MATHEMATICAL Female f	MATHEMATICAL Female % below	TSWE Male f	TSWE Male % below	TSWE Female f	TSWE Female % below
800	1	100.0	0	100.0	13	99.9	0	100.0	0	100.0	0	100.0
790	1	100.0	1	100.0	17	99.9	2	100.0	0	100.0	0	100.0
780	1	100.0	1	100.0	17	99.8	5	100.0	0	100.0	0	100.0
770	2	100.0	0	100.0	30	99.6	3	100.0	0	100.0	0	100.0
760	4	100.0	7	100.0	48	99.4	6	99.9	0	100.0	0	100.0
750	10	99.9	8	99.9	78	99.2	15	99.9	0	100.0	0	100.0
740	14	99.8	25	99.8	92	98.7	15	99.8	0	100.0	0	100.0
730	16	99.8	14	99.7	151	98.0	32	99.6	0	100.0	0	100.0
720	13	99.7	9	99.7	120	97.5	24	99.5	0	100.0	0	100.0
710	49	99.5	33	99.5	134	96.9	38	99.3	0	100.0	0	100.0
700	11	99.4	21	99.4	154	96.2	40	99.2	0	100.0	0	100.0
690	61	99.2	48	99.2	148	95.5	44	98.9	0	100.0	0	100.0
680	49	98.9	32	99.1	145	94.6	70	98.6	0	100.0	0	100.0
670	90	98.5	75	98.7	197	93.7	66	98.3	0	100.0	0	100.0
660	74	98.2	49	98.5	256	92.5	71	98.0	0	100.0	0	100.0
650	66	97.9	57	98.2	219	91.5	82	97.6	0	100.0	0	100.0
640	165	97.1	150	97.5	234	90.5	89	97.2	0	100.0	0	100.0
630	91	96.7	47	97.3	293	89.1	122	96.6	0	100.0	0	100.0
620	224	95.7	191	96.4	305	87.7	134	96.0	0	100.0	0	100.0
610	126	95.1	120	95.8	345	86.1	146	95.3	0	100.0	0	100.0
600	250	94.0	217	94.8	655	82.1	390	93.4	642	97.1	720	96.6
590	153	93.3	122	94.2	429	81.2	221	92.4	375	95.2	393	94.8
580	368	91.6	317	92.7	397	79.3	246	91.2	0	95.3	0	94.8
570	201	90.7	196	91.8	435	77.4	267	90.0	477	93.2	468	92.5
560	192	89.8	152	91.1	376	75.6	265	88.7	565	90.6	599	89.7
550	478	87.6	434	89.0	513	73.3	324	87.2	68	90.3	68	89.4
540	269	86.4	251	87.8	1080	68.3	692	83.9	626	87.4	678	84.2
530	536	83.9	524	85.4	500	66.0	346	82.3	706	84.2	725	82.8
520	341	82.3	318	83.9	497	63.8	391	80.5	759	80.7	710	79.4
510	672	79.3	662	80.8	568	61.2	484	78.2	732	77.3	864	75.4
500	379	77.5	362	79.1	601	58.4	488	75.9	144	76.7	151	74.7
490	737	74.2	729	75.6	589	55.7	496	73.5	767	73.2	802	70.9
480	467	72.0	428	73.6	1109	50.6	1014	68.7	808	69.5	844	66.9
470	483	69.8	442	71.5	619	47.8	542	66.2	896	65.4	925	62.5
460	967	65.4	895	67.3	625	44.9	567	63.5	852	61.4	883	58.4
450	505	63.1	487	65.0	582	42.3	539	61.0	239	60.4	221	57.3
440	1082	58.1	1055	60.0	514	39.9	545	58.4	764	56.9	765	52.7
430	569	55.5	543	57.5	1096	34.9	1292	52.3	824	53.1	804	49.9
420	1009	50.9	949	53.0	541	32.4	619	49.4	737	49.7	802	46.1
410	587	48.2	570	50.3	460	30.3	523	46.9	777	46.1	822	42.3
400	545	45.7	611	47.4	471	28.2	562	44.2	256	44.8	252	41.1
390	1016	41.0	972	42.8	483	26.0	657	41.1	743	41.4	738	37.6
380	553	38.5	567	40.2	494	23.7	603	38.3	761	37.9	731	34.2
370	1128	33.2	1021	35.3	842	19.8	1035	33.4	725	34.6	665	31.0
360	554	30.8	511	32.9	449	17.8	536	30.9	0	34.6	0	31.0
350	860	26.9	813	29.1	446	15.7	615	28.0	711	31.3	646	28.0
340	477	24.7	475	26.9	419	13.8	611	25.1	311	29.9	265	26.7
330	908	20.5	891	22.7	319	12.4	480	22.8	617	27.1	579	24.0
320	358	18.9	290	21.3	657	9.3	1070	17.8	617	24.3	574	21.3
310	444	16.8	358	19.6	347	7.8	606	14.9	579	21.6	552	18.7
300	703	13.6	775	16.0	271	6.5	527	12.5	550	19.1	463	16.5
290	387	11.9	334	14.4	232	5.5	410	10.5	265	17.8	214	15.5
280	567	9.3	545	11.8	249	4.3	473	8.3	449	15.8	390	13.7
270	261	8.1	296	10.4	256	3.1	457	6.1	445	13.8	353	11.8
260	504	5.8	582	7.7	336	1.6	619	3.2	418	11.8	365	10.1
250	150	5.1	205	6.7	112	1.1	203	2.3	360	10.2	327	8.5
240	360	3.4	449	4.6	97	0.6	182	1.4	187	9.3	157	7.8
230	162	2.7	188	3.7	69	0.3	136	0.8	310	7.9	237	6.7
220	128	2.1	183	2.8	34	0.2	67	0.4	283	6.6	228	5.6
210	162	1.3	220	1.8	29	0.0	71	0.1	270	5.4	226	4.5
200	275	0.0	382	0.0	0	0.0	24	0.0	1175	0.0	963	0.0
	21,835		21,209		21,835		21,209		21,835		21,209	
Mean	415.1		407.9		472.9		420.1		405.1		414.1	
S.D.	107.4		108.1		118.8		106.9		111.2		109.2	

the female study group) tend to be given to scores between 240 and 550 on the verbal scale, scores between 260 and 540 on the mathematical scale, and scores between 30 and 60 on the TSWE scale (a relatively large weight is also assigned to 20).

Results

SAT Verbal Results

Table 2 contains listings of four indices described earlier, P_f, $\hat{P}_f$, D_f, and $RMWSD_f$, and the observed percent correct in the male base group, P_m, for the 85 verbal items of Form ZSA5. In addition, it includes the means and standard deviations of these five indices displayed by item type.

The first row of the summary portion of Table 2 contains statistics based on all 85 verbal items. Note that mean P_f and mean $\hat{P}_f$ are equal to two decimals. The difference between mean D_f (.00) and mean $RMWSD_f$ (.05) is attributed to the fact that $RMWSD_f$, unlike D_f, is an unsigned index of discrepancy that weights and sums any squared differences between P_f and $\hat{P}_f$ regardless of which value is larger and thus prevents cancellation of positive and negative differences. On the other hand, the signed index D_f expresses the amount by which total differences in one direction exceed total differences in the other direction.

The next row in Table 2 displays the means and standard deviations of the five indices computed on the vocabulary items only. Again, mean P_f and mean $\hat{P}_f$ are nearly equal. Both discrepancy indices are also small. The vocabulary items can be divided still further into antonym items and analogies items. Mean percent correct on these item types are even less related to scaled scores than previous item groupings, and so differences in mean percent correct are

Table 2

Listing of Item Difficulty and Discrepancy Indices and Summary Statistics for Verbal Items from SAT Form ZSA5

ITEM #	ITEM TYPE	P_f % CORRECT	$\hat{P}_f$ EST % CORRECT	D_f DIFF % CORRECT	RMSD	P_b % CORRECT BASE GROUP
1	ANTONYM	0.6905	0.6507	0.0398	0.0508	0.6702
2	ANTONYM	0.7143	0.6864	0.0279	0.0406	0.7045
3	ANTONYM	0.8098	0.7632	0.0166	0.0442	0.8030
4	ANTONYM	0.7062	0.6928	0.0134	0.0293	0.7104
5	ANTONYM	0.7666	0.7066	0.0600	0.0693	0.7223
6	ANTONYM	0.6853	0.7401	-0.0511	0.0615	0.7531
7	ANTONYM	0.5125	0.5658	-0.0533	0.0661	0.5777
8	ANTONYM	0.4818	0.4787	0.0031	0.0273	0.4926
9	ANTONYM	0.5796	0.5156	0.0640	0.0697	0.5276
10	ANTONYM	0.2223	0.1910	0.0213	0.0547	0.1944
11	ANTONYM	0.3449	0.3452	-0.0002	0.0335	0.3531
12	ANTONYM	0.2924	0.2664	0.0261	0.0401	0.2745
13	ANTONYM	0.3009	0.2439	0.0570	0.0754	0.2517
14	ANTONYM	0.0786	0.0801	-0.0015	0.0435	0.0801
15	ANTONYM	0.1383	0.1256	0.0127	0.0524	0.1268
16	SENT COM	0.7825	0.8373	-0.0547	0.0716	0.8523
17	SENT COM	0.6569	0.7435	-0.0866	0.0603	0.7572
18	SENT COM	0.6951	0.6774	0.0176	0.0309	0.6919
19	SENT COM	0.6916	0.7010	-0.0094	0.0307	0.7147
20	SENT COM	0.4032	0.4955	-0.0923	0.1054	0.5047
21	READ COM	0.5361	0.5158	0.0203	0.0391	0.5287
22	READ COM	0.6518	0.6866	0.0052	0.0362	0.6993
23	READ COM	0.6321	0.6423	-0.0102	0.0330	0.6543
24	READ COM	0.5616	0.5701	-0.0085	0.0316	0.5801
25	READ COM	0.2607	0.2360	0.0248	0.0471	0.2420
26	READ COM	0.0917	0.1075	-0.0158	0.0381	0.1111
27	READ COM	0.2301	0.2399	-0.0097	0.0428	0.2482
28	READ COM	0.1142	0.1422	-0.0280	0.0503	0.1450
29	READ COM	0.1307	0.1910	-0.0603	0.0760	0.1974
30	READ COM	0.1568	0.1395	0.0173	0.0385	0.1430
31	SENT COM	0.8455	0.8345	0.0111	0.0403	0.8457
32	SENT COM	0.4822	0.5068	-0.0246	0.0409	0.5182
33	SENT COM	0.4469	0.3788	0.0681	0.0788	0.3899
34	SENT COM	0.3363	0.3259	0.0104	0.0365	0.3359
35	SENT COM	0.1308	0.1186	0.0122	0.0275	0.1255
36	ANALOGY	0.7601	0.8354	-0.0752	0.0843	0.8443
37	ANALOGY	0.7435	0.7905	-0.0429	0.0542	0.7149
38	ANALOGY	0.6261	0.6401	-0.0140	0.0293	0.6541
39	ANALOGY	0.5276	0.5117	0.0159	0.0253	0.5274
40	ANALOGY	0.4669	0.4515	0.0154	0.0343	0.4643
41	ANALOGY	0.3405	0.3647	-0.0242	0.0381	0.3756
42	ANALOGY	0.2447	0.2112	0.0334	0.0466	0.2193
43	ANALOGY	0.1367	0.1656	-0.0289	0.0480	0.1733
44	ANALOGY	0.1498	0.1979	-0.0483	0.0691	0.2047
45	ANALOGY	0.0671	0.0884	-0.0212	0.0334	0.0912
46	ANTONYM	0.8831	0.8607	0.0225	0.0327	0.8724
47	ANTONYM	0.8085	0.7684	0.0401	0.0520	0.7823
48	ANTONYM	0.7160	0.8034	-0.0874	0.0975	0.8173
49	ANTONYM	0.7037	0.6639	0.0398	0.0485	0.6773
50	ANTONYM	0.4443	0.4015	0.0428	0.0561	0.4153
51	ANTONYM	0.4750	0.4631	0.0119	0.0393	0.4760
52	ANTONYM	0.4575	0.3872	0.0703	0.0832	0.3980
53	ANTONYM	0.3569	0.3755	-0.0186	0.0327	0.3866
54	ANTONYM	0.1282	0.1732	-0.0449	0.0714	0.1781
55	ANTONYM	0.1079	0.1038	0.0041	0.0310	0.1075
56	SENT COM	0.7701	0.8155	-0.0454	0.0606	0.8295
57	SENT COM	0.6744	0.6244	0.0500	0.0661	0.6392
58	SENT COM	0.7046	0.7139	-0.0094	0.0260	0.7203
59	SENT COM	0.4893	0.4439	0.0453	0.0565	0.4584
60	SENT COM	0.2218	0.2051	0.0167	0.0424	0.2125

Table 2 (continued)

ITEM #	ITEM TYPE	P_f % CORRECT	$\hat{P}_f$ EST % CORRECT	D_f DIFF % CORRECT	P-BSD	P_b % CORRECT BASE GROUP
61	ANALOGY	0.8360	0.8276	0.0083	0.0255	0.8426
62	ANALOGY	0.8191	0.8402	-0.0211	0.0346	0.8510
63	ANALOGY	0.6295	0.7990	-0.1695	0.1762	0.8111
64	ANALOGY	0.6078	0.6377	-0.0299	0.0441	0.6487
65	ANALOGY	0.4652	0.5289	-0.0637	0.0775	0.5365
66	ANALOGY	0.1355	0.1297	0.0058	0.0336	0.1249
67	ANALOGY	0.2206	0.2361	-0.0155	0.0436	0.2451
68	ANALOGY	0.1475	0.1279	0.0196	0.0495	0.1252
69	ANALOGY	0.1489	0.1415	0.0073	0.0359	0.1448
70	ANALOGY	0.1593	0.1403	0.0190	0.0373	0.1458
71	READ COM	0.4790	0.4389	0.0401	0.0548	0.4527
72	READ COM	0.6052	0.5567	0.0485	0.0575	0.5702
73	READ COM	0.5180	0.4744	0.0436	0.0588	0.4866
74	READ COM	0.7328	0.6934	0.0394	0.0490	0.7097
75	READ COM	0.2214	0.2284	-0.0070	0.0312	0.2349
76	READ COM	0.2065	0.1950	0.0114	0.0353	0.2007
77	READ COM	0.6245	0.5612	0.0633	0.0732	0.5749
78	READ COM	0.4006	0.3759	0.0247	0.0389	0.3904
79	READ COM	0.6430	0.5724	0.0706	0.0787	0.5866
80	READ COM	0.4851	0.4322	0.0529	0.0620	0.4437
81	READ COM	0.1825	0.1783	0.0042	0.0317	0.1850
82	READ COM	0.2443	0.2751	-0.0308	0.0440	0.2849
83	READ COM	0.2099	0.2414	-0.0315	0.0427	0.2520
84	READ COM	0.2261	0.2494	-0.0234	0.0398	0.2602
85	READ COM	0.1607	0.1764	-0.0157	0.0353	0.1813

Item Type	No. of Items	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
All Verbal	85	.45	.24	.45	.24	.00	.04	.05	.02	.46	.25
Vocabulary	45	.46	.26	.46	.26	.00	.04	.05	.03	.47	.27
Antonyms	25	.50	.25	.49	.25	.01	.04	.05	.02	.50	.25
Analogies	20	.41	.26	.43	.27	-.02	.05	.05	.03	.44	.28
Reading	40	.44	.23	.44	.22	.00	.04	.05	.02	.45	.23
Sentence Comp.	15	.56	.21	.56	.22	-.00	.04	.05	.02	.57	.22
Reading Passages	25	.37	.21	.36	.19	.01	.03	.05	.01	.37	.19

more likely to appear among antonyms or analogies item type groupings than in the vocabulary items as a whole or the entire verbal test.

The next two rows of Table 2 list the means and standard deviations of the five indices across the antonyms and analogies item types, respectively. The values of $RMWSD_f$ are still of approximately the same size as before. The magnitudes of D_f are slightly larger than before, yet still small in an absolute sense.

The statistics for reading, the other section in the verbal test, and the two item types that compose it, sentence completion and reading comprehension, and the corresponding statistics from their items are posted in the last three rows of Table 2. None of these indices exhibit disconcerting amounts of unexpected differential item performance.

Even if the overall level of unexpected differential item performance in a set of items is tolerable, there may be some small number of items which exhibit substantial unexpected differential item performance that is not readily detectable from the means and standard deviations of discrepancy indices such as $RMWSD_f$ and D_f. For an item level analysis, careful examination of the frequency distribution of a discrepancy index such as $RMWSD_f$ can be informative. A combination numerical/pictorial display of the frequency distribution of the RMWSD index on all verbal items grouped by subscore and by item type is presented in Figure 5. The floating histogram in Figure 5 is a clear presentation of the $RMWSD_f$ index that can be used to identify individual items that exhibit unusually high amounts of unexpected differential item performance. Note how the single analogy item with a RMSWD of .18 clearly stands out in this figure.

Figure 5

Numerical and Pictorial Display of Frequencies of Root Mean Weighted Squared Differences(RMWSD) Between the Conditional Probabilities of Success for Female and Male Candidates on Verbal Items from Form ZSA5 Administered in December 1977

Numerical Frequencies Grouped by Item Type

Floating Histograms by Item Type

VERB	VOCAB	ANTM	ANAL	READ	SNCP	RDCP	Values of RMWSD	Vocabulary		Reading	
								ANTM	ANAL	SNCP	RDCP
							.20				
							.19				
1	1		1				.18		A		
							.17				
							.16				
							.15				
							.14				
							.13				
							.12				
1				1	1		.11			S	
1	1	1					.10	0			
							.09				
7	4	2	2	3	1	2	.08	00	AA	S	RR
8	5	4	1	3	2	1	.07	0000	A	SS	R
8	2	2		6	3	3	.06	00		SSS	RRR
13	9	5	4	4		4	.05	00000	AAAA		RRRR
25	10	5	5	15	4	11	.04	00000	AAAAA	SSSS	RRRRRRRRRR
21	13	6	7	8	4	4	.03	000000	AAAAAAA	SSSS	RRRR
							.02				
							.01				
							.00				
.04	.03	.03	.03	.04	.035	.04	Mode				
.05	.05	.05	.05	.05	.05	.05	Mean				
.02	.03	.02	.03	.02	.02	.01	S.D.				

Legend:

Item Type	No. of Items	Abbreviations
Verbal Score	85	(VERB)
Vocabulary Subscore	45	(VOCAB)
Antonyms	25	(ANTM) (O)
Analogies	20	(ANAL) (A)
Reading Subscore	40	(READ)
Sentence Completion	15	(SNCP) (S)
Reading Comprehension	25	(RDCP) (R)

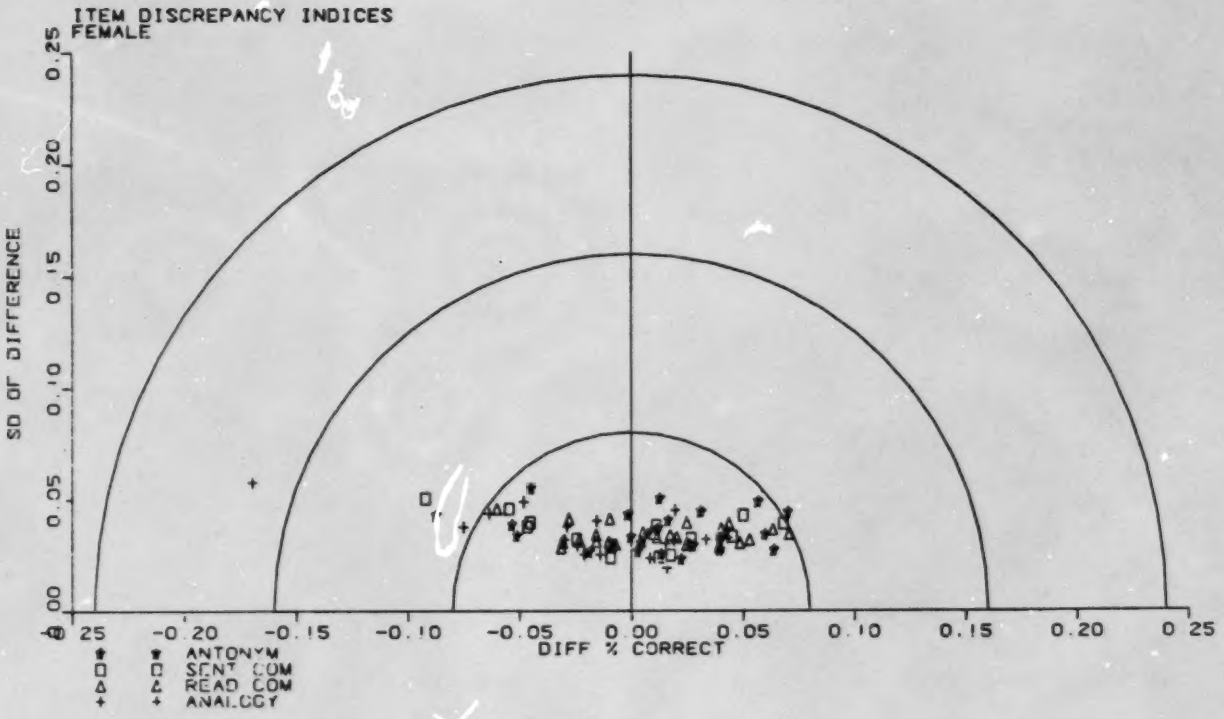
An alternative pictorial representation of the distribution of this index that conveys even more information is given in Figure 6. In this figure, where each item type is denoted by a different symbol, the $RMWSD_f$ for an item is represented by the length of the line from the origin to the point representing that item. To supply a frame of reference, three arcs of equal $RMWSD_f$ are drawn on the plot for the values .08, .16 and .24. Items falling within the smallest arc exhibit a fairly typical amount of $RMWSD_f$. Items falling between the smallest and middle arc should be examined more closely. Items falling outside the middle arc are very unusual and clearly exhibit a large amount of unexpected differential item performance.

As described earlier, the $RMWSD_f$ for each item can be expressed as the square root of two additive components, the square of a constant directional discrepancy, which is D_f, and an index of residual crossover, i.e., a sum of weighted squared differences in conditional probabilities after correction for the constant difference, which is referred to as the variance of the weighted differences. (See equation (4).) Projection of each point in Figure 6 on the horizontal axis yields the D_f, the difference between P_f and $\hat{P}_f$, for that item. Projection of that same point on the vertical axis yields the standard deviation of the weighted differences, the index of residual crossover. Hence, the location of each point in Figure 6 indicates not only the degree of unexpected differential item performance ($RMWSD_f$), but also the extent to which that $RMWSD_f$ is due to a constant difference between the P_f and $\hat{P}_f$ curves (and the direction of that difference: D_f), and the extent to which the item exhibits residual crossover, the height of the point above the horizontal axis.

The analogy item depicted in Figures 3 and 4 is the only verbal item which falls outside the second arc of Figure 6. It is also the item in Figure 5 that

Figure 6

Plot of Root Mean Weighted Squared Differences (RMWSD[a]) Between the Conditional Probabilities of Success for Male and Female Candidates on Verbal Items from SAT Form ZSA5



[a]RMWSD equals the distance from the origin to the point representing the item. Projection of each point on the horizontal axis yields the difference between P_f and $\hat{P}_f$, D_f, for that item. Projection of each point on the vertical axis yields the standard deviation of the weighted differences, an index of residual crossover.

is off by itself in the floating histogram at the top where it has a $RMWSD_f$ of .1792. Clearly this index indicates a highly undesirable amount of unexpected differential item performance for this analogy item.

In Figure 6, the analogy item outside the second arc is just above .05 on the vertical axis and at approximately -.17 on the horizontal axis. Hence, this item is exhibiting little residual crossover, and a very sizeable amount of constant difference. Examination of Figure 4 corroborates these observations. This analogy item exhibits a substantial constant amount of unexpected differential item performance.

In contrast to this item, most of the items fall within the first arc, which indicates that most of the items, 80 out of 85 in fact, exhibit acceptable levels of unexpected differential item performance. Of the four that fall between the inner and middle arcs, an antonym item that has a positive D_f and an analogy item with a negative D_f are close enough to the inner arc to be considered as exhibiting acceptable levels of unexpected differential item performance. The remaining two items, a sentence completion item and an antonym, however, merit some careful examination. Like the analogy item outside the middle arc, these two items have negative D_f values, which indicate that female candidates perform poorer than expected on these items.

On the analogy item that lies outside the middle arc, female candidates performed far worse than expected: $P_f = .63$ vs. $\hat{P}_f = .80$. Inspection of the content of this particular analogy item revealed potential content bias against female candidates, as it required some knowledge of hunting and fishing, two traditionally male-oriented recreational activities.

On the sentence completion item, female candidates performed somewhat lower than expected: $P_f = .40$ vs. $\hat{P}_f = .50$. Inspection of this item itself revealed that the subject matter of the item, nuclear power politics, might be something that males traditionally have shown more interest in than females. It is not apparent, however, why this particular subject matter should affect the performance of female candidates on this item.

Finally, on the antonym item, female candidates performed below expectation: $P_f = .72$ vs. $\hat{P}_f = .80$. Examination of item content, however, provided no plausible explanation for this difference.

In sum, this analysis of the 85 verbal items on Form ZSA5 uncovered only one item that exhibited a substantial amount of unexpected differential item performance that probably could be attributed to item content. Only two other items exhibited enough unexpected differential item performance to merit examination. Most of the 85 verbal items exhibited little unexpected differential item performance for female candidates.

SAT-Mathematical Results

Table 3 contains listings of the five indices, P_f, $\hat{P}_f$, D_f, $RMWSD_f$, and P_m for the 60 mathematics items on Form ZSA5. In addition, these indices are summarized by item type at the bottom of this table. The first row at the bottom of Table 3 contains means and standard deviations based on 59 mathematics items. One math item was excluded from this analysis because the percent of female candidates responding correctly to the item was less than .05.

Unlike verbal test results, mean P_f (.42) for female candidates and mean P_m (.51) for male candidates are very different, reflective of the difference between the mathematical ability distributions for males and females, and

Table 3

Listing of Item Difficulty and Discrepancy Indices and Summary Statistics for Mathematical Items from SAT Form ZSA5

ITEM #	ITEM TYPE	P_f % CORRECT	$\hat{P}_f$ EST % CORRECT	D_f DIFF % CORRECT	PNWSD	P_b % CORRECT BASE GROUP
1	REG MATH	0.7340	0.7312	0.0028	0.0276	0.7569
2	REG MATH	0.5502	0.5259	-0.0756	0.0863	0.7261
3	REG MATH	0.5377	0.5656	-0.0279	0.0420	0.6879
4	REG MATH	0.6456	0.6168	0.0289	0.0405	0.7029
5	REG MATH	0.4965	0.5200	-0.0335	0.0480	0.6359
6	REG MATH	0.6741	0.6171	0.0569	0.0675	0.7134
7	REG MATH	0.6765	0.6759	0.0006	0.0254	0.7774
8	REG MATH	0.4966	0.5430	-0.0465	0.0603	0.6629
9	REG MATH	0.6124	0.5421	0.0703	0.0826	0.6616
10	REG MATH	0.4258	0.4270	-0.0012	0.0346	0.5209
11	REG MATH	0.4683	0.4782	-0.0099	0.0345	0.5721
12	REG MATH	0.4787	0.4511	0.0276	0.0423	0.5671
13	REG MATH	0.5990	0.5581	0.0408	0.0584	0.6468
14	REG MATH	0.4117	0.3949	0.0168	0.0338	0.4898
15	REG MATH	0.4247	0.4589	-0.0302	0.0448	0.5727
16	REG MATH	0.2297	0.2595	0.0302	0.0553	0.3816
17	REG MATH	0.1420	0.1672	-0.0251	0.0395	0.2513
18	REG MATH	0.1471	0.1841	-0.0370	0.0454	0.2570
19	REG MATH	0.2546	0.2403	0.0143	0.0449	0.3223
20	REG MATH	0.0970	0.0928	0.0042	0.0214	0.1354
21	REG MATH	0.1286	0.1727	-0.0441	0.0570	0.2425
22	REG MATH	0.1087	0.1114	-0.0027	0.0244	0.1661
23	REG MATH	0.0848	0.0821	0.0027	0.0195	0.1226
24	REG MATH	0.0936	0.1121	-0.0184	0.0338	0.1593
25	REG MATH	0.0552	0.0598	-0.0046	0.0224	0.0957
26	REG MATH	0.8201	0.8150	0.0051	0.0300	0.8776
27	REG MATH	0.5307	0.5903	-0.0596	0.0791	0.6888
28	REG MATH	0.8174	0.8055	0.0120	0.0248	0.8686
29	REG MATH	0.6529	0.6951	-0.0422	0.0546	0.7818
30	REG MATH	0.6272	0.6561	-0.0288	0.0470	0.7315
31	REG MATH	0.5911	0.5856	0.0055	0.0269	0.6818
32	REG MATH	0.5748	0.5603	0.0145	0.0330	0.6374
33	QUANTCMP	0.6634	0.6639	-0.0005	0.0265	0.7609
34	QUANTCMP	0.6594	0.5790	0.0804	0.0884	0.6537
35	QUANTCMP	0.6869	0.6638	0.0231	0.0375	0.7345
36	QUANTCMP	0.6090	0.5860	0.0230	0.0413	0.6971
37	QUANTCMP	0.7183	0.7740	-0.0557	0.0665	0.8489
38	QUANTCMP	0.4279	0.4152	0.0127	0.0308	0.5375
39	QUANTCMP	0.5955	0.5402	0.0553	0.0693	0.6248
40	QUANTCMP	0.5782	0.5963	-0.0180	0.0323	0.6987
41	QUANTCMP	0.5825	0.6035	-0.0210	0.0390	0.6922
42	QUANTCMP	0.5061	0.5364	-0.0303	0.0443	0.6377
43	QUANTCMP	0.5303	0.5281	0.0022	0.0296	0.6306
44	QUANTCMP	0.4110	0.4040	0.0070	0.0289	0.4987
45	QUANTCMP	0.4650	0.4596	0.0004	0.0312	0.5781
46	QUANTCMP	0.3747	0.3862	-0.0114	0.0258	0.4980
47	QUANTCMP	0.1866	0.1943	-0.0077	0.0369	0.2842
48	QUANTCMP	0.2551	0.2483	0.0108	0.0330	0.3284
49	QUANTCMP	0.3335	0.3175	0.0160	0.0425	0.3499
50	QUANTCMP	0.1461	0.1775	-0.0313	0.0417	0.2363
51	QUANTCMP	0.1615	0.1986	-0.0370	0.0752	0.2635
52	QUANTCMP	0.2050	0.2149	-0.0099	0.0289	0.2704
53	REG MATH	0.5176	0.4523	0.0652	0.0721	0.5515
54	REG MATH	0.1826	0.1767	0.0059	0.0209	0.2898
55	REG MATH	0.3214	0.3803	-0.0589	0.0701	0.4690
56	REG MATH	0.0875	0.1042	-0.0167	0.0258	0.1705
57	REG MATH	0.1543	0.1513	0.0027	0.0354	0.2099
58	REG MATH	0.0851	0.1068	-0.0177	0.0294	0.1468
59	REG MATH	0.0782	0.0872	-0.0089	0.0264	0.1131
60[a]	REG MATH	0.0490	0.0532	-0.0042	0.0177	0.0788

Item Type	No. of Items	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
Mathematical	59	.42	.22	.42	.21	.00	.03	.04	.02	.51	.23
Quantitative Comparison	20	.45	.18	.45	.18	.00	.03	.04	.02	.54	.19
Regular Math	39	.40	.24	.41	.23	.00	.03	.04	.02	.49	.24

[a]Excluded from summary statistics because P_f less than .05.

illustrative of the need to correct for this difference prior to comparing male and female item performance. Note that mean $\hat{P}_f$ (.42), in contrast to mean P_m, is very close to mean P_f, demonstrating the effectiveness of the standardization procedure in this regard. Both D_f and $RMWSD_f$ have very low means, indicating little overall difference, as expected, between the sexes on the items.

The next row of Table 3 displays the means and standard deviations of the five indices computed on the 20 quantitative comparison items. Female candidates' mean percent correct is extremely close to their estimated mean (i.e., mean D_f = .00). The mean value of $RMWSD_f$ is only .04.

The last row of Table 3 presents the data for 39 regular math type items. Item #60 was excluded from the analysis because the female candidates' percent correct on this item was less than .05. These means and standard deviations suggest that little unexpected differential item performance is present.

Figures 7 and 8 contain pictorial and numerical displays of the discrepancy indices for both quantitative comparison and regular mathematics item types. Neither the floating histogram in Figure 7 nor the plot in Figure 8 reveal any items that exhibit the substantial degree of unexpected differential item performance observed for the one analogy item in the verbal test. Only two items, in fact, fall outside the inner arc in Figure 8. Female candidates performed better than expected on one item, but more poorly than expected on the other item. The plots of male and female conditional percent corrects and the difference plot for the former item are given in Figures 9 and 10, respectively, while Figures 11 and 12 are the corresponding plots for the latter item. Note that Figures 9 and 11 appear to be mirror images of each other, with female candidates slightly exceeding male candidates in Figure 9, while the reverse

Figure 7

Numerical and Pictorial Display of Frequencies of Root Mean Weighted Squared Differences (RMWSD) Between the Conditional Probabilities of Success for Female and Male Candidates on Mathematical Items from Form ZSA5 Administered in December 1977

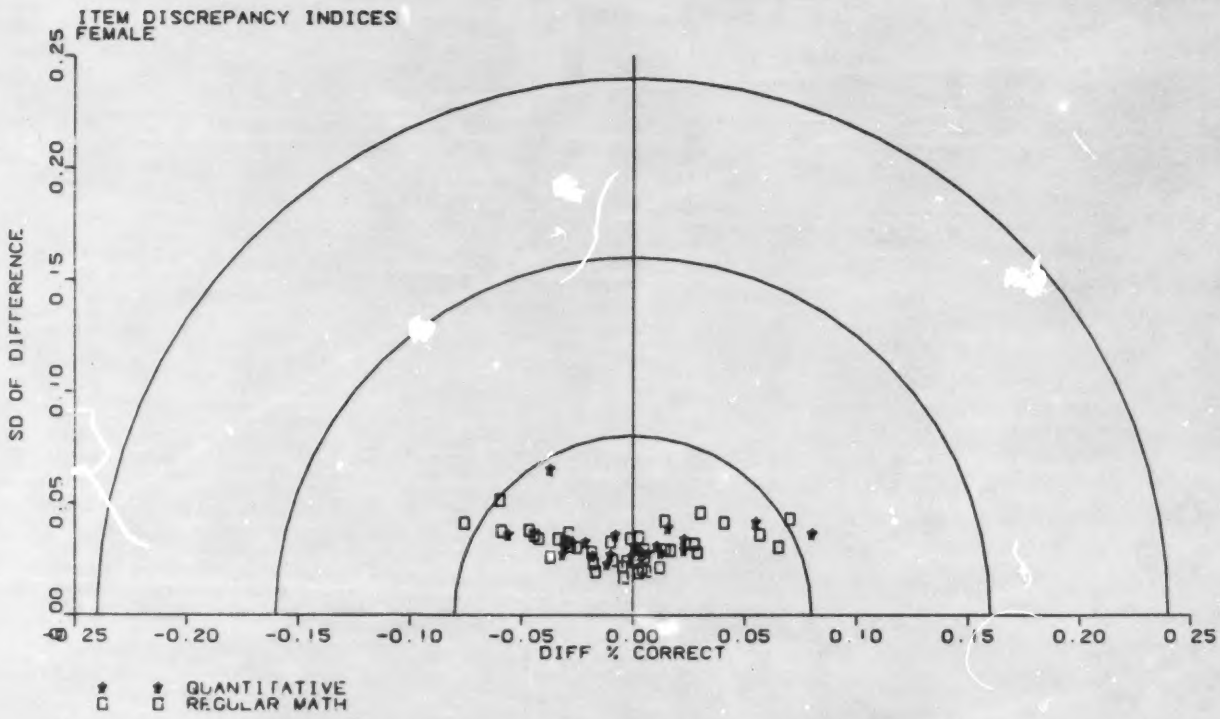
Numerical Frequencies Grouped by Item Type			Values of MWAD	Floating Histograms by Item Type	
Mathematical Score	Regular Mathematics	Quantitative Comparison		Regular Mathematics	Quantitative Comparison
			.20		
			.19		
			.18		
			.17		
			.16		
			.15		
			.14		
			.13		
			.12		
			.11		
			.10		
2	1	1	.09	R	Q
3	2	1	.08	RR	Q
5	3	2	.07	RRR	QQ
5	4	1	.06	RRRR	Q
4	4		.05	RRRR	
13	7	6	.04	RRRRRRR	QQQQQQ
21	12	9	.03	RRRRRRRRRRR	QQQQQQQQQ
7	7		.02	RRRRRRR	
			.01		
			.00		
.03	.03	.03	Mode		
.04	.04	.04	Mean		
.02	.02	.02	S.D.		

Legend:

Item Type	No. of Items	Abbreviations
Mathematical Score	60	
Regular Mathematics	40	(R)
Quantitative Comparison	20	(Q)

Figure 8

Plot of Root Mean Weighted Squared Differences (RMWSD[a]) Between the Conditional Probabilities of Success for Male and Female Candidates on Mathematics Items from SAT Form ZSA5



[a]RMWSD equals the distance from the origin to the point representing the item. Projection of each point on the horizontal axis yields the difference between P_f and $\hat{P}_f$, D_f, for that item. Projection of each point on the vertical axis yields the standard deviation of the weighted differences, an index of residual crossover.

Conditional Probability of Successful Performance for Both Males and Females on Two Math Items from SAT Form ZSA5

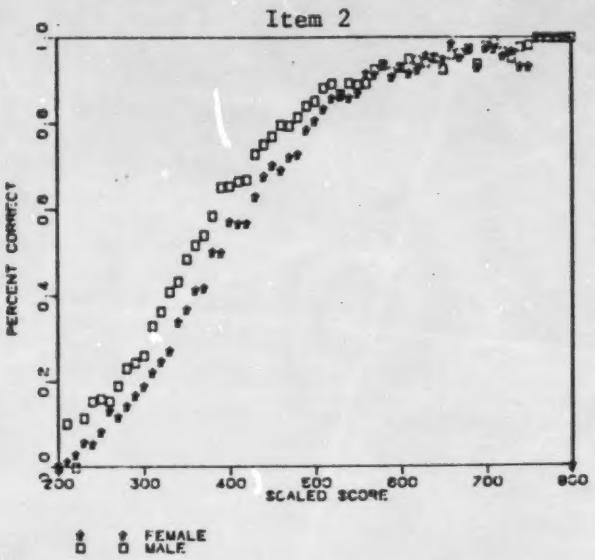


Figure 9

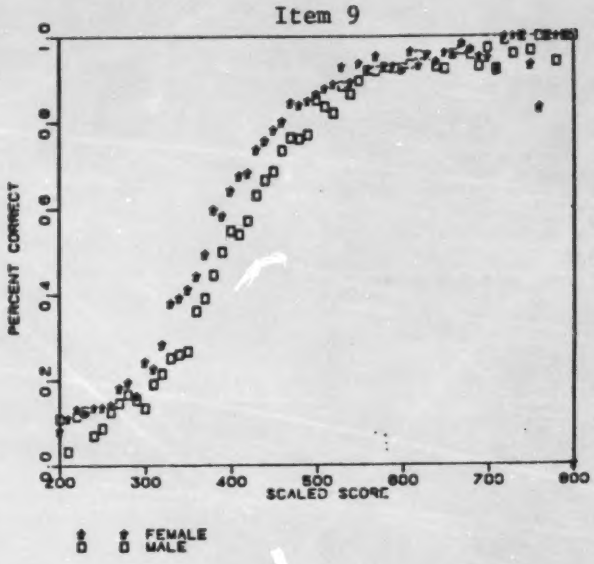


Figure 11

Difference Plot of Two Math Items from SAT Form ZSA5

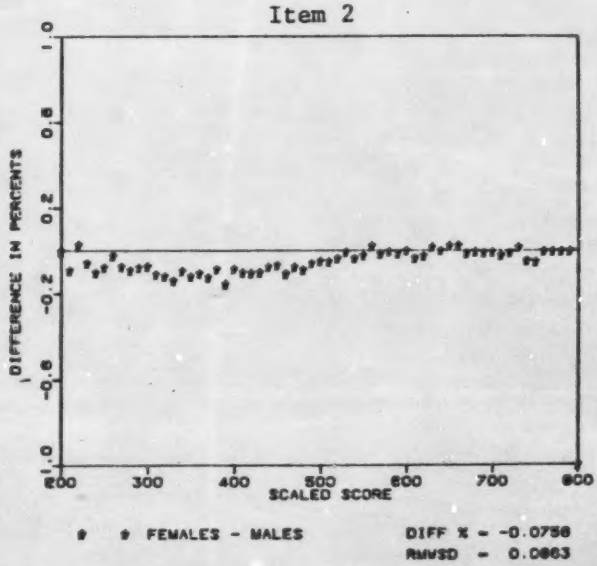


Figure 10

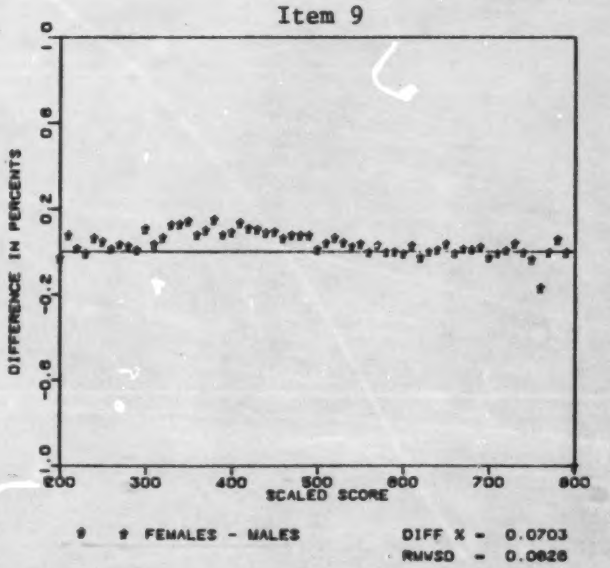


Figure 12

occurs in Figure 11. Both figures exhibit fairly constant differences, but in opposite directions. Examination of the content of these two items provided no apparent explanation for these differences. Hence, it appears that all mathematics items on Form ZSA5 are relatively free from unexpected differential item performance for females, despite the fact that the mean scaled score for female candidates was approximately one-half a standard deviation lower than the male candidate mean scaled score. The standardization procedure effectively adjusted for this difference in overall performance.

TSWE Total Test and Item Type Results

Table 4 contains a listing of the five indices, P_f, $\hat{P}_f$, D_f, $RMWSD_f$, and P_m, discussed in preceding sections, for the 50 TSWE items on Form E11. In addition, these indices are summarized by item type at the bottom of the table. The first row at the bottom of Table 4 contains means and standard deviations based on all 50 TSWE items, and the next two rows contain the same information for the 35 usage type items and the 15 sentence correction items, respectively. Estimated percent correct ($\hat{P}_f$) means for the female candidates are very close to actual (P_f) means across both item types combined and separately. The mean values of RMWSD are similar to those observed for the mathematical items. No mean differences appear large enough to warrant further consideration.

Figures 13 and 14 contain pictorial and numerical displays of the discrepancy indices for all TSWE items on Form E11. Inspection of these figures reveals that only two usage items exhibit any substantial amounts of unexpected differential item performance. Performance on these items is depicted in greater detail in Figures 15-18. The female candidates performed better than expected (P_f = .59 vs. $\hat{P}_f$ = .50) on the item displayed in Figures 15 and 16.

Table 4

Listing of Item Difficulty and Discrepancy Indices and Summary Statistics for TSWE Items from Form E11

ITEM #	ITEM TYPE	P_f % CORRECT	$\hat{P}_f$ EST % CORRECT	D_f DIFF % CORRECT	RMWSD	P_b % CORRECT BASE GROUP
1	USAGE	0.9214	0.9249	-0.0035	0.0152	0.9181
2	USAGE	0.8060	0.8162	-0.0102	0.0299	0.8020
3	USAGE	0.7846	0.7982	-0.0136	0.0317	0.7826
4	USAGE	0.5853	0.4971	0.0881	0.0974	0.4856
5	USAGE	0.6952	0.6724	0.0228	0.0338	0.6593
6	USAGE	0.7619	0.7906	-0.0287	0.0373	0.7756
7	USAGE	0.8918	0.8794	0.0125	0.0240	0.8690
8	USAGE	0.3236	0.3308	-0.0072	0.0372	0.3224
9	USAGE	0.6329	0.5927	0.0402	0.0543	0.5819
10	USAGE	0.5449	0.5731	-0.0282	0.0415	0.5551
11	USAGE	0.7765	0.7756	0.0009	0.0207	0.7605
12	USAGE	0.7916	0.7845	0.0071	0.0253	0.7710
13	USAGE	0.8479	0.8362	0.0117	0.0229	0.8248
14	USAGE	0.8081	0.8042	0.0040	0.0263	0.7936
15	USAGE	0.5096	0.5935	-0.0839	0.0928	0.5781
16	USAGE	0.3768	0.3743	0.0025	0.0299	0.3622
17	USAGE	0.5772	0.5998	-0.0226	0.0352	0.5799
18	USAGE	0.6233	0.6202	0.0031	0.0271	0.6067
19	USAGE	0.4848	0.4953	-0.0105	0.0319	0.4819
20	USAGE	0.5658	0.5808	-0.0151	0.0340	0.5638
21	USAGE	0.8381	0.8129	0.0252	0.0389	0.8065
22	USAGE	0.4546	0.5342	-0.0797	0.0483	0.5155
23	USAGE	0.5391	0.5261	0.0030	0.0294	0.5185
24	USAGE	0.8650	0.8530	0.0160	0.0250	0.8423
25	USAGE	0.4400	0.4269	0.0131	0.0353	0.4118
26	SENT COR	0.9206	0.9152	0.0054	0.0186	0.9073
27	SENT COR	0.7735	0.7405	0.0330	0.0379	0.7268
28	SENT COR	0.7275	0.7333	-0.0058	0.0386	0.7213
29	SENT COR	0.8726	0.8615	0.0112	0.0289	0.8511
30	SENT COR	0.7676	0.7977	-0.0101	0.0263	0.7858
31	SENT COR	0.6979	0.7188	-0.0188	0.0310	0.7009
32	SENT COR	0.6423	0.6426	-0.0003	0.0179	0.6218
33	SENT COR	0.6411	0.6320	0.0091	0.0262	0.6163
34	SENT COR	0.4612	0.4625	-0.0013	0.0281	0.4490
35	SENT COR	0.6497	0.6931	-0.0434	0.0588	0.6771
36	SENT COR	0.7107	0.6935	0.0173	0.0323	0.6742
37	SENT COR	0.5867	0.6039	-0.0172	0.0234	0.5833
38	SENT COR	0.5393	0.5359	0.0035	0.0280	0.5164
39	SENT COR	0.3624	0.4282	-0.0657	0.0712	0.4079
40	SENT COR	0.4403	0.4348	0.0055	0.0275	0.4164
41	USAGE	0.8426	0.8303	0.0123	0.0213	0.8123
42	USAGE	0.4461	0.4553	-0.0092	0.0245	0.4386
43	USAGE	0.7661	0.7506	0.0155	0.0259	0.7295
44	USAGE	0.5008	0.4527	0.0482	0.0548	0.4272
45	USAGE	0.6383	0.6265	0.0119	0.0368	0.6087
46	USAGE	0.6521	0.6326	0.0195	0.0387	0.6132
47	USAGE	0.5829	0.5809	0.0019	0.0295	0.5596
48	USAGE	0.6651	0.6209	0.0441	0.0511	0.5973
49	USAGE	0.6139	0.5732	0.0407	0.0543	0.5543
50	USAGE	0.3151	0.2989	0.0162	0.0295	0.2839

Item Type	No. of Items	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
TSWE	50	.65	.16	.64	.16	.00	.03	.04	.02	.63	.16
Usage	35	.64	.16	.64	.16	.01	.03	.04	.02	.62	.16
Sentence Correction	15	.65	.15	.66	.14	-.01	.02	.03	.01	.64	.15

Figure 13

Numerical and Pictorial Display of Frequencies of Root Mean Weighted Squared Differences(RMWSD) Between the Conditional Probabilities of Success for Female and Male Candidates on TSWE Items from Form ZSA5/E11 Administered in December 1977

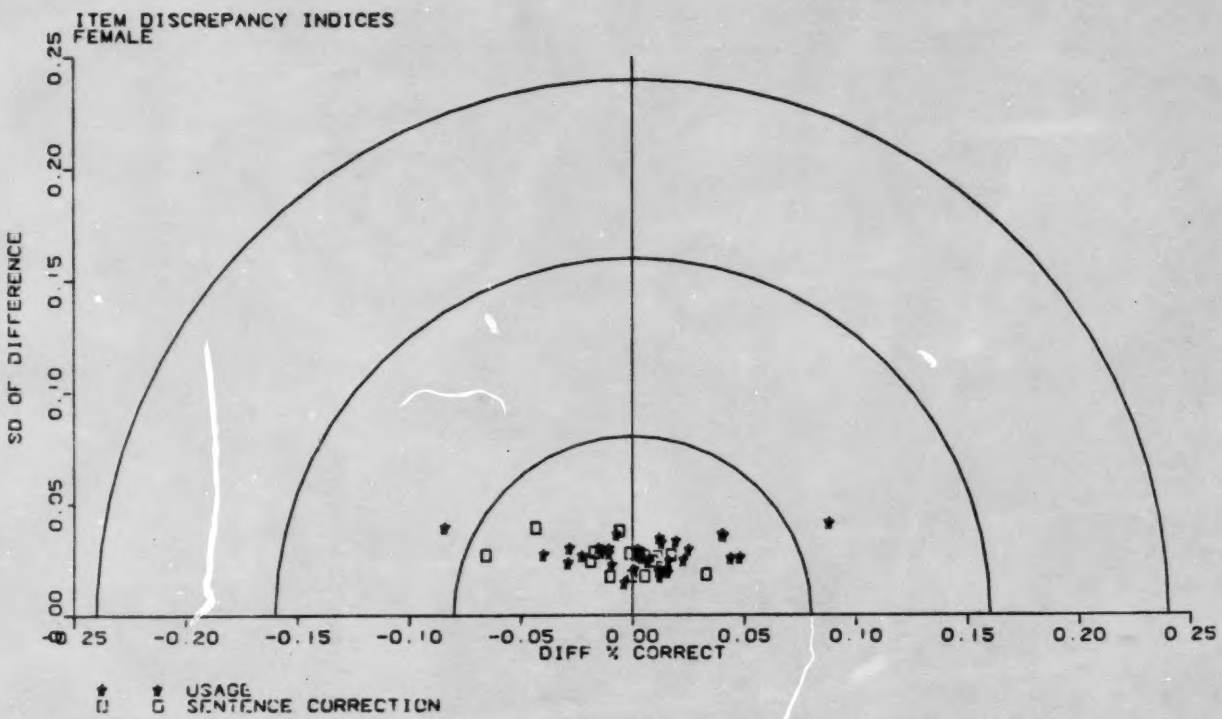
Numerical Frequencies Grouped by Item Type			Values of RMWSD	Floating Histograms by Item Type	
TSWE Score	Usage	Sentence Correction		Usage	Sentence Correction
			.20		
			.19		
			.18		
			.17		
			.16		
			.15		
			.14		
			.13		
			.12		
			.11		
1	1		.10	U	
1	1		.09	U	
			.08		
1		1	.07		C
1		1	.06		C
5	5		.05	UUUUU	
10	8	2	.04	UUUUUUUU	CC
21	13	8	.03	UUUUUUUUUUUUU	CCCCCCCC
10	7	3	.02	UUUUUUU	CCC
			.01		
			.00		
.03	.03	.03	Mode		
.04	.04	.03	Mean		
.02	.02	.01	S.D.		

Legend:

Item Type	No. of Items	Abbreviations
Test of Standard Written English Score	50	TSWE
Usage	35	(U)
Sentence Correction	15	(C)

Figure 14

Plot of Root Mean Weighted Squared Differences (RMWSD[a]) Between the Conditional Probabilities of Success for Male and Female Candidates on TSWE Items from Form ZSA5/E11



[a]RMWSD equals the distance from the origin to the point representing the item. Projection of each point on the horizontal axis yields the difference between P_f and $\bar{P}_f$, D_f, for that item. Projection of each point on the vertical axis yields the standard deviation of the weighted differences, an index of residual crossover.

Conditional Probability of Successful Item Performance
for Both Males and Females on Two
TSWE Items from Form ZSA5/E11

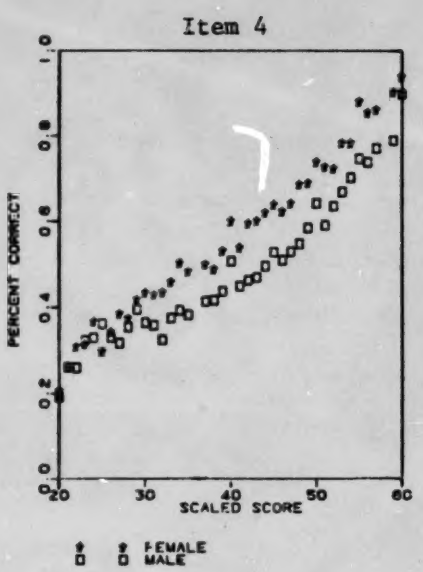


Figure 15

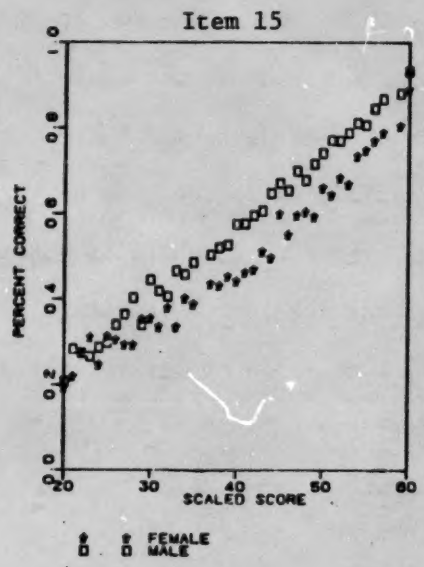


Figure 17

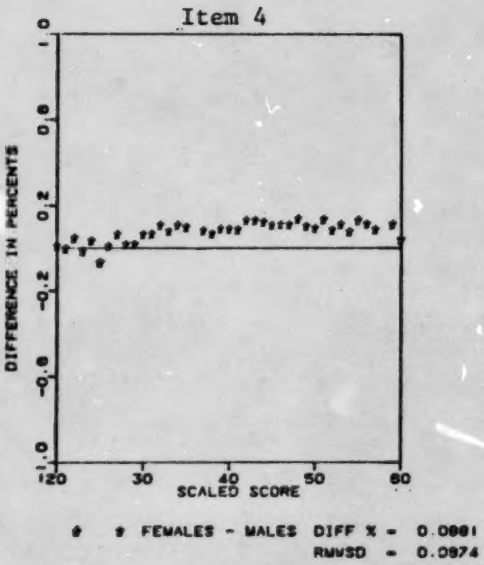


Figure 16

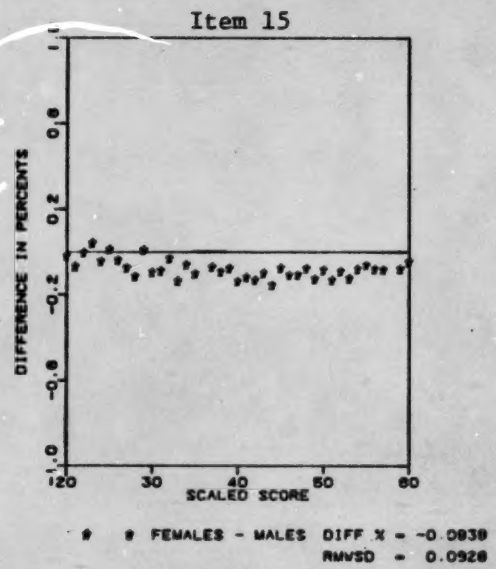


Figure 18

Most of this difference is constant across levels of scaled score. On the item displayed in Figures 17 and 18, the female candidates did not perform as well as expected, (P_f = .51, $\hat{P}_f$ = .59). Again, most of the difference is in one direction. Note that these two items appear to cancel each other out.

Examination of the content of these two items revealed that the item on which females performed better than expected concerns a woman in a professional occupation, while the item on which females fell short of expectation deals with World War II, which is generally considered an area that males study more than females. However, these content differences do not appear to be sufficient explanations for the discrepancies in the observed and expected performance of female candidates on these items.

Summary

This report was the first in a series of investigations seeking to uncover evidence relating to the presence or absence of unexpected differential item performance on operational SAT/TSWE items across different candidate subpopulations of the SAT/TSWE test-taking population via the statistical method of standardization. The use of standardization enables one to control for differences in subpopulation ability. Standardization is a reasonable procedure for controlling for differences in ability, provided the control variable is a reasonable measure of ability, as is total scaled score.

Examination of summary statistics for discrepancy indices at the item type level revealed that there was little evidence of systematic unexpected differential item performance on either the SAT-M or TSWE tests. On the verbal test, the analogy items exhibited a mean D_f which suggested systematic unexpected

differential item performance that favored the male candidates. Elimination of the one analogy item which exhibited very substantial unexpected differential item performance reduces the mean D_f for analogy items by half when that item is included in the set, i.e., from -.02 to -.01, suggesting that with the exception of that one item, the analogy items, as a set, exhibit little unexpected differential item performance.

In contrast to previous investigations of item fairness (see review by Dorans, 1982), this investigation of differential item performance identified very few items out of a total of 195 items as needing careful review for possible content bias. Of these only one exhibited a clearly unacceptable degree of unexpected differential item performance that could be attributed to content bias.

Since this is the first application of the standardization approach to studies of unexpected differential item performance, future applications are bound to involve modifications of the method as employed here. Certain modifications are very likely to occur. For example, different candidate subpopulations will be studied and, as a consequence, the range of scaled scores studied may be curtailed. A variation of the standardization procedure that can be used with small samples may be employed. For some studies, the focus may be shifted away from breakdowns by item type towards breakdowns by content, where feasible. In short, the methodology will be refined and adapted to meet the requirements of future applications.

References

Alderman, D. L. and Holland P. W. Item performance across native language groups on the Test of English as a Foreign Language (RR81-16), Princeton, NJ: Educational Testing Service, 1981.

Berk, R. A. (Ed.) Handbook of methods for detecting test bias. Baltimore, MD: Johns Hopkins Press, 1982.

Carlton, S. T., and Marco, G. L. Methods used by test publishers to "debias" standardized tests: Educational Testing Service. In R. A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore, MD: Johns Hopkins Press, 1982.

Cook, L., and Nutkowitz, I. Test analysis: College Board Scholastic Aptitude Test December 1977 Administration ZSA5 (SR-79-63). Princeton, NJ: Educational Testing Service, 1979.

Donlon, T. F. The SAT in a diverse society: Fairness and sensitivity. The College Board Review, No. 122 (Winter 1981-82), 16-21, 30-32.

Dorans, N. J. Technical review of item fairness studies: 1975-1979 (SR-82-90). Princeton, NJ: Educational Testing Service, 1982.

Hecht, L. W., and Swineford, F. Item analysis at Educational Testing Service, Princeton, NJ: Educational Testing Service, 1981.

Shepard, L., Camilli, G., and Averill, M. Comparison of six procedures for detecting test item bias using both internal and external ability criteria. Journal of Educational Statistics, 1981, 6, 317-375.

Stern, J. Test analysis: College Board Test of Standard Written English June 1976 Administration E11(YSA3), Princeton, NJ: Educational Testing Service, 1977.

Walker, R. C. A reader's guide to test analysis reports. Princeton, NJ: Educational Testing Service, 1981.

Appendix

THE STANDARDIZATION APPROACH TO ASSESSING UNEXPECTED DIFFERENTIAL ITEM PERFORMANCE

Since the standardization approach to assessing unexpected differential item performance represents a new application of an old technique to an important concern in applied testing, the approach will be presented in detail in this appendix. First, the rationale for standardization will be discussed. Then, the particular application of standardization will be described. In the process of describing this approach to assessing unexpected differential item performance, several terms and concepts will be defined. The goals of this appendix are:

(1) to convey the simplicity and generality of the standardization approach, and

(2) to illustrate its application to the assessment of unexpected differential item performance.

The Need for Standardization

Standardization is a statistical technique that enables one to compare two populations of individuals with respect to some variable of interest while controlling for differences on some other variable that is related to the variable of interest. The best way to convey the meaning and importance of standardization is to illustrate what may occur when standardization is not performed when it should be. Simpson's paradox is the designation for a paradoxical situation in which a population with a higher overall incidence of some variable than a second population actually has a lower incidence of that variable than the second population when comparisons of that variable are conditioned on some other variable. Simpson's paradox (Wagner, 1982) can be used to illustrate the importance of standardization.

Consider the following illustration. Table 1 contains a statistical description of the performance of two hypothetical groups, A and B, on an item. Group A is composed of 100,000 candidates, while Group B is composed of 1,000 candidates. In the body of the table, the performance of the two groups on the item is summarized at the far right under the column heading overall performance. Here we note that 60,000 of the 100,000 members of Group A answered the item correctly, while 500 of the 1,000 members of Group B answered the item correctly. Since the 60% for Group A exceeds the 50% for Group B, we might conclude that this particular item favors Group A over Group B. Such an interpretation, however, would be in error because it ignores important information about the two groups that is contained in the rest of the table, namely that Group A is more able than Group B.

To the left of the overall performance column in Table 1 are five columns of numbers that describe the performance on the item of subgroups of A and B that are classified into five mutually exclusive performance levels, L1-L5. As is evident in the %-Correct rows of the table, L1 is the least able subgroup, L5 is the most able, and L2, L3 and L4 are ordered from low to high in terms of performance on the control variable. At each ability level, members of Group A are as able as members of Group B. Thus, the 35,000 members of Group A at L4 are as able as the 150 members of Group B at L4.

The numbers in the first and fifth rows of the table identify the number of individuals in Groups A and B, respectively, at each of the performance levels. These numbers inform us that overall Group A is more able than Group B with most of Group A at levels L4 and L5 and most of Group B at L2 and L3. This substantial difference in overall ability between Groups A and B affects the summary infor-

Table 1

Performance of Two Groups of Different Ability on an Item that Favors the Lower Ability Group

	Ability Level					
	L1	L2	L3	L4	L5	Overall Performance
Group A						
No. of Individuals	5000	15000	25000	35000	20000	100000
% at Level	.05	.15	.25	.35	.20	
Answer Correct	500	4500	12500	24500	18000	60000
% Correct	.1	.3	.5	.7	.9	.6
Group B						
No. of Individuals	200	350	250	150	50	1000
% at Level	.20	.35	.25	.15	.05	
Answer Correct	40	140	150	120	50	500
% Correct	.2	.4	.6	.8	1.0	.5

mation portrayed in the overall performance column, which had led us to conclude that the item favored Group A over Group B.

A closer examination of all the information in Table 1, however, leads us to conclude that the item, in fact, favors Group B over Group A. The evidence for this conclusion is contained in the fourth and eighth rows of Table 1, which contain the percent correct for each of the five ability levels in groups A and B, respectively. Note that at each ability level, a larger percentage of Group B members answer the item correctly than do Group A members of comparable ability. This analysis, conditioned on ability level, indicates that this item favors Group B over Group A because the probability of successful performance on the item is .1 higher for Group B than Group A at each of the five ability levels. Simpson's paradox refers to the fact that the analysis conditioned on ability level contradicts the analysis based on a simple comparison of overall performance of the two groups on the item, i.e., the analysis based on the data in the overall performance column of Table 1.

Standardization with respect to ability level removes the paradox in the item performance analyses by producing a simple total group comparison, like that based on the overall performance column, which is not confounded by differences in group ability. Standardization accomplishes this goal by using the same standard ability distribution for both groups.

Definitions

In the balance of this appendix, the following definitions will be employed to designate various subgroups and variables used by the standardization approach to the assessment of unexpected differential item performance:

Variables. There are two types of variables: study and control. The study variable is the variable of interest, while the control variable is a variable that is related to the study variable and which must be controlled while making comparisons of the study variable. In the example under consideration, performance on the item expressed as percent correct is the study variable, while ability level is the control variable. Since percent correct is related to ability level, the latter must be controlled for during comparisons of the former.

Groups. There are three types of groups: study, standardization, and base. The study group, as the phrase implies, is the group under study. In any given investigation, there are as many potential study groups as there are potential subgroups in a population. In actuality, certain subgroups, e.g. Blacks, are more likely to be study groups because of concerns about the relevance of tests for these subgroups.

The standardization group supplies the ability distributions used by the standardization approach. In any comparison of two groups, three possible standardization groups immediately suggest themselves: either of the two groups or a composite of the two groups. While all three of these groups are based on actual data, the standardization approach is not limited to standardization groups based on actual data. A hypothetical ability distribution constructed to suit some desiderata could be used as the standardization group.

The base group supplies the model for the data to the standardization process. The model for the data expresses the study variable as a function of the control variable. In assessing unexpected differential item performance, the model is the expected performance on the item conditioned on ability, i.e., the expected probability of successful performance on the item given ability

level. As in the case of the study group, there are as many potential base groups as there are potential subgroups. A subgroup cannot be both the study group and the base group in the same analysis, however. To achieve a stable model for data, the base group should be as large as possible. To avoid part-total group contaminations, the base group should be independent of the various study groups in an investigation.

In investigations of unexpected differential item performance, the model for the data can be empirical or theoretical. An example of an empirical model in an investigation of unexpected differential item performance in a Black study group would be the conditional percent correct in a white base group. If an adjustment of percent correct for not reached, omits and number wrong served as the data for the study group, an empirical model of the data would be the comparable adjusted percent correct observed in the base group. Further discussion of adjusted percent correct is reserved for the mathematical formalization presented latter in this appendix.

The various models of item response theory (Lord, 1980) are examples of theoretical models for the data. This appendix is limited to empirical models for the data.

Mathematical Formalization

The mathematical formulation of the standardization approach to assessing unexpected differential item performance can be described in several stages, each of which focuses on a different component. These components are:

I. Observed Study Group Data

   A. Basic Data

   B. Derived Data to be Modelled

II. The Model for the Data

III. Definition of the Standardization Group

IV. Statistical Indices of Unexpected Differential Item Performance

Observed Study Group Data

In the balance of this appendix, the following indices will be employed:

- g is the subscript for subgroup and ranges from 1 to G, where G is the number of subgroups;
- s is the subscript for scaled score or ability level and ranges from 1 to S, where S is the number of scaled score levels. For SAT-V and SAT-M, S is 61; for TSWE, S is 41;
- r is a response type indicator for which
  1 = correct response
  2 = incorrect response
  3 = omit
  4 = not reached.

Basic Data. The basic data are counts, N_{gsr}, i.e., the number (frequency) of people in subgroup g at ability level s who gave response type r to the item. For example, N_{gs1} is the number of people in g at ability level s who responded correctly to the item, while N_{gs3} is the number of people in g at ability level s who omitted the item. If we let "+" represent a simple unweighted sum, then N_{gs+} is the number of people in g at s. In addition, $N_{gs+} - N_{gs4}$ is the number of people in g at s who reached the item.

Derived Data to be Modelled. Some variation of percent correct are the data to be modelled for unexpected differential item performance. Simple percent correct at ability level s in subgroup g is defined as

(1) $$P_{gs} = N_{gs1} / N_{gs+} .$$

An alternative percent correct involves a correction for not reached,

(2) $$P_{gs}(NR) = N_{gs1} / (N_{gs+} - N_{gs4}) .$$

Yet another "adjusted" percent correct entails an adjustment for guessing,

(3) $$P_{gs}(GA) = (N_{gs1} - N_{gs2}/(k-1)) / N_{gs+} ,$$

where k is the number of options in the multiple choice question. Choice of "percent correct" depends on the purposes of the investigation. Various choices, such as (1) - (3) above, can be obtained as a special case of a general formula for the data,

(4) $$P_{gs}(W_r) = \frac{\sum_{r=1}^{4} N_{gsr} * w_r^t}{\sum_{r=1}^{4} N_{gsr} * w_r^b}$$

where w_r^t is the rth element in the vector of weights $\underline{W}_r^t$ applied to N_{gsr} to obtain the numerator of $P_{gs}(W_r)$, while w_r^b is the rth element in the vector of weights $\underline{W}_r^b$ applied to N_{gsr} to obtain the denominator of $P_{gs}(W_r)$. For equations (1) to (3) above, the corresponding weight vectors, $\underline{W}_r^t$ and $\underline{W}_r^b$ are:

Equation	$\underline{W}_r^t$	$\underline{W}_r^b$
	R, W, O, NR	R, W, O, NR
(1)	(1, 0, 0, 0)	(1, 1, 1, 1)
(2)	(1, 0, 0, 0)	(1, 1, 1, 0)
(3)	(1,-1/(k-1), 0, 0)	(1, 1, 1, 1)

Choice of $\underline{W}_r^t$ and $\underline{W}_r^b$ for use in (4) determines the data $P_{gs}(W_r)$ to be modelled. In the example in Table 1, simple percent correct, equation (1), was used to obtain the data to be modelled. Dividing the numbers in the third and seventh rows by the numbers in the first and fifth rows, respectively, provides the simple percent corrects contained in the fourth and eighth rows, respectively, of Table 1. For example, the .5 (P_{AL3}) for group A at score level L3 is obtained by dividing 12,500 (N_{AL31}) by 25,000 (N_{AL3+}).

The Model for the Data

The data are defined as the percent correct for the study group. For an empirical model, the model for the data is simply the same percent correct for the base group. Both the data and the empirical model for the data are obtained via equation (4). For the data, the subscript g refers to the study group. Likewise, for the model, the subscript g refers to the base group.

When the data base is sufficiently large, as in the case with the SAT, it is often sensible to use the largest subgroup as the base group. In that case the model for the data can be obtained via a straightforward application of equation (4). In the hypothetical example depicted in Table 1, the base group model values for simple percent correct data are simply the observed percent correct data for group A, which are listed in the fourth row.

Definition of the Standardization Group

The standardization group supplies the standard ability distributions used by the standardization approach. Any of the G subgroups can be used as the standardization group. Since the standard ability distribution serves as a weighting function, it is advisable to use each study group as its own standard-

ization group thereby using a weighting function that mirrors the relative frequency at each score level in the study group.

Formalizing the role of the standard ability distribution in the standardization process illustrates how it serves as a weighting function. As the phrase might imply, "unexpected differential item performance" focuses on unexpected differences in item performances. Controlling for differences in subgroup ability through standardization, enables us to label as unexpected any difference between actual and expected item performance. For subgroups composed of equally able members, there should be no differences in item performance. For the SAT and TSWE, reported scaled scores are highly reliable measures of the developed abilities assessed by that testing instrument. It is therefore reasonable to presume that individuals at the same scaled score ability level across subgroups should have the same probability of successful performance on the item. Hence unexpected differential item performance focuses on differences in item performance at fixed score levels. For SAT-V and SAT-M, there are 61 reported score levels, and for TSWE, there are 41 reported score levels. Standardization affords us with a simple way of summarizing unexpected differences in each item performance across score levels. For both SAT-V and SAT-M, it enables us to reduce 61 potential differences to two summary indices without the confounding effects due to differences in group ability. For TSWE, 41 potential differences are reduced to two summary indices.

Statistical Indices of Unexpected Differential Item Performance

At each score level s, in group g, we have the difference,

$$D_{gs} = P_{gs} - \hat{P}_{gs} , \tag{5}$$

where P_{gs} is observed data defined in (4) using the study groups counts, N_{gsr}, and $\hat{P}_{gs}$ is the model for the data defined via (4) using the base groups counts. In equation (5), D_{gs} is a conditional difference between the data and the model. Let W_{gs} be the standardization group weighting function for study group g. A sensible weighting function contains the relative frequencies of scaled score s in study group g, i.e.,

$$W_{gs} = N_{gs+} / N_{g++} , \tag{6}$$

where N_{gs+} is the number of individuals in group g at score level s and N_{g++} is the number of individuals in group g across all s score levels.

Applying each W_{gs} to its corresponding conditional difference and summing across score levels yields a mean weighted difference,

$$D_g = \sum_{s=1}^{S} W_{gs} D_{gs} \tag{7}$$

an overall difference between the data and the model for percent correct. This difference is one index of unexpected differential item performance supplied by standardization with respect to ability. A second index is the mean weighted squared difference,

$$MWSD = \sum_{s=1}^{S} W_{gs} D_{gs}^2 \tag{8}$$

which can be rewritten as

$$MWSD = \sum_{s=1}^{S} W_{gs} D_{gs} D_{gs} \tag{9}$$

which implies that each difference is weighted by itself as well as by the weighting function associated with the standardization group. The square root

of MWSD is also an index of discrepancy, RMWSD, that is on a scale that is comparable to D_g.

To illustrate the standardization process, let us return to the data in Table 1. Suppose Group B were the study group, chosen as such because its lower ability level led critics of testing to believe that test items were biased against Group B. Since there are 100,000 individuals in group A, it was chosen as the base group. Since we are primarily interested in study group B, its ability distribution supplies us with a natural weighting function. Hence, the data, model and weighting function are:

	P_{BS}	$\hat{P}_{BS} = P_{AS}$	W_{BS}
L1:	$\frac{40}{200} = .20$	$\frac{500}{5000} = .10$	$\frac{200}{1000} = .20$
L2:	$\frac{140}{350} = .40$	$\frac{4500}{15000} = .30$	$\frac{350}{1000} = .35$
L3:	$\frac{150}{250} = .60$	$\frac{12500}{25000} = .50$	$\frac{250}{1000} = .25$
L4:	$\frac{120}{150} = .80$	$\frac{24500}{35000} = .70$	$\frac{150}{1000} = .15$
L5:	$\frac{50}{50} = 1.0$	$\frac{18000}{20000} = .90$	$\frac{50}{1000} = .05$
			$\Sigma = 1.0$

Note that, as with all weighting functions, $\Sigma W_{BS} = 1.0$. Using the information above, we obtain

	P_{BS}	$\hat{P}_{BS} = P_{AS}$	D_{BS}	W_{BS}	$W_{BS}D_{BS}$	$W_{BS}D_{BS}^2$
L1:	.2	.1	.1	.20	.020	.0020
L2:	.4	.3	.1	.35	.035	.0035
L3:	.6	.5	.1	.25	.025	.0025
L4:	.8	.7	.1	.15	.015	.0015
L5:	1.0	.9	.1	.05	.005	.0005
				Σ =1.0	Σ =.1	Σ =.01

The last row above reveals that D_B = .1 and $MWSD_B$ = .01 when Group A is the base group. Note that, $D_B^2 = MWSD_B$, which indicates that all the sum of squared differences are due to the constant difference of .1 observed at each score level.

Contrasting Standardization With Other Approaches

The assessment of unexpected differential item performance is an important concern in applied testing. As such it has attracted much attention, e.g., Berk's (1982) Handbook of Methods for Detecting Test Bias. From the title of Berk's volume one might infer that several methods for bias detection exist, and the contents of the volume confirm this inference. The intent of this closing section is to place the standardization approach within the context of the methods included in the Berk volume.

Scheuneman (1981) makes a distinction between two general types of item bias definitions: definitions related to an item-by-group interaction, e.g., Angoff's (Angoff and Ford, 1973) transformed item difficulty approach, and definitions that involve conditioning on ability, e.g., item response theory

approaches (Lord, 1980). Unexpected differential item performance is clearly a definition involving conditioning on ability. The standardization approach to assessing unexpected differential item performance is most akin to item response theory methods.

In item response theory approaches, parameterized item-ability regressions, or item response functions, for different subgroups are computed and compared. In the standardization approach, unparameterized item-test regressions are compared. While the parametric nature of the item response methods are more elegant, the particular model (e.g., one-parameter), may not fit the data and the lack of fit might be misconstrued as bias. In contrast, unparameterized item-test regressions will not suffer from model fit problems. Like any method that uses an internal criterion, however, unparameterized item-test regressions are subject to bothersome item-total contaminations.

While the standardization approach is more akin to parametric item response theory methods, it shares some of the simplicity of the transformed item difficulty or delta-plot method. It too results in "transformed" item difficulties, namely the predicted p-values obtained from applying the marginal ability distribution of the standardization group to the base group conditional item success curves. These predicted p-values are the item difficulties one would expect if both the base group and the study group had ability distributions like that of the standardization group. These predicted difficulties should be identical because ability has been directly controlled for through standardization. Any substantial deviation from identity could be construed as evidence of unexpected differential item performance, evidence stated in the simple metric of proportion answering an item correctly.

Reference

Angoff, W. H., and Ford, S. F. Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 1973, 10, 95-106.

Berk, R. A. (Ed.) Handbook of methods for detecting test bias. Baltimore, MD: Johns Hopkins University Press, 1982.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum, 1980.

Scheuneman, J. D. A new look at bias in aptitude tests. In P. Merrifield (Ed.), New Directions for Testing and Measurement: Measuring human abilities, No. 12. San Francisco: Jossey-Bass, 1981.

Wagner, C. H. Simpson's paradox in real life. The American Statistician, 1982, 36 (1), 46-47.